

A logical account of trust in information sources

Robert Demolombe, Emiliano Lorini

Institut de Recherche en Informatique de Toulouse (IRIT), France
robert.demolombe@orange.fr
lorini@irit.fr

Abstract. We present a model of trust that integrates in trust definition: the truster's motivation (his goal), the action that allows the trustee to reach this goal, and the trustee's ability and intention to do this action. This model is formalized in modal logic and it is applied to the particular domain of trust on information sources. In this context trust may be derived, in particular, from truster's beliefs about some trustee's epistemic properties: sincerity, competence, vigilance, cooperativity, validity and completeness.

1 Introduction

Trust on information sources plays an role in many areas of interactions between agents, in particular when some information sources are software agents. A typical example is in the field of stock and bond market where trust has a strong influence on agents' behavior when they decide to buy, or to sale, a specific kind of stocks. To take such decisions agents have several types of information sources to know expectations about the future evolution of the value of stocks. They may be banks, companies, consultants or others, and the agents may believe that some of these sources have a good competence but are not necessarily sincere, others are reluctant to inform about bad news, others are competent but are not necessarily informed at the right moment...

We see that reasoning about so complex situations requires clear definitions of the relevant concepts, and safe reasoning rules to infer consequences from communication actions that have been performed, or that have not been performed. The objective of this paper is to propose a formal framework for this purpose. However, it is out of the scope of this work to propose a model to give support for agents' trust about such or such information source property like, for example, statistics based on observations, or reputation evaluation.

We start in section 2 with the presentation of the logical framework which is used in the following sections. Then, a general definition of trust is presented in section 3. The section 4 is devoted to the formal definitions of agents' epistemic properties like: sincerity, competence, and others. These properties are used in section 5 to apply general definitions of trust in the field of trust on information sources. A comparison with other related works is presented in section 6 and the conclusion shows directions for future extensions.

2 A logic for trust reasoning

The logic \mathcal{L} we use to formalize the relevant concepts involved in our model of social trust is a multimodal logic which combines the expressiveness of a simple dynamic logic [11] with the expressiveness of a logic of mental attitudes [5, 19] and obligations [1, 3]. The syntactic primitives of the logic \mathcal{L} are the following:

- a nonempty finite set of agents $AGT = \{i, j, \dots\}$;
- a nonempty finite set of atomic actions $AT = \{a, b, \dots\}$;
- a set of atomic formulas $\Pi = \{p, q, \dots\}$.

We denote with LIT the set of literals include all atomic formulas and their negations, that is, $LIT = \{p, \neg p \mid p \in \Pi\}$. We denote with P, Q, \dots the elements in LIT . We also introduce specific actions of the form $inf_j(P)$ denoting the action of informing agent j about P . We call them informative actions. The set $INFO$ of informative actions is defined as follows:

- $INFO = \{inf_j(P) \mid j \in AGT, P \in LIT\}$.

The set ACT of complex actions is given by the union of the set of atomic actions and the set of informative actions, that is:

- $ACT = AT \cup INFO$.

We denote with symbols α, β, \dots the elements in ACT .

The language of \mathcal{L} is the set of formulas defined by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid After_{i:\alpha}\phi \mid Does_{i:\alpha}\phi \mid Bel_i\phi \mid Goal_i\phi$$

where p ranges over Π , α ranges over ACT and i ranges over AGT .

The operators of our logic have the following intuitive meaning. $Bel_i\phi$: the agent i believes that ϕ ; $After_{i:\alpha}\phi$: after agent i does α , it is the case that ϕ ($After_{i:\alpha}\perp$ is read: agent i cannot do action α); $Does_{i:\alpha}\phi$: agent i is going to do α and ϕ will be true afterward ($Does_{i:\alpha}\top$ is read: agent i is going to do α); $Goal_i\phi$: the agent i wants that ϕ holds. The following abbreviations are given:

$$Can_i(\alpha) \stackrel{\text{def}}{=} \neg After_{i:\alpha}\perp;$$

$$Int_i(\alpha) \stackrel{\text{def}}{=} Goal_i Does_{i:\alpha}\top;$$

$$Inf_{i,j}(p) \stackrel{\text{def}}{=} Does_{i:inf_j(p)}\top.$$

We write $Can_i(\alpha)$ as an abbreviation of $\neg After_{i:\alpha}\perp$ in order to make explicit the fact that $\neg After_{i:\alpha}\perp$ stands for: agent i can do action α (i.e. i has the capacity/ability to do α). $Int_i(\alpha)$ stands for: agent i intends to do α . Finally, $Inf_{i,j}(p)$ stands for: agent i informs agent j about p .

Models of our logic are tuples $M = \langle W, R, D, B, G, V \rangle$ where:

- W is a non empty set of possible worlds or states.
- R is a collection of binary relations $R_{i:\alpha}$ on W , one for every couple $i:\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w', w) \in R_{i:\alpha}$ then w' is a world which can be reached from world w through the occurrence of agent i 's action α .

- D is a collection of binary relations $D_{i:\alpha}$ on W , one for every couple $i:\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w', w) \in D_{i:\alpha}$ then w' is the *next* world of w which will be reached from w through the occurrence of agent i 's action α .
- B is a collection of binary relations B_i on W , one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w', w) \in B_i$ then w' is a world which is compatible with agent i 's beliefs at world w .
- G is a collection of binary relations G_i on W , one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w', w) \in G_i$ then w' is a world which is compatible with agent i 's goals at world w .
- $V : \Pi \rightarrow 2^W$ is a valuation function.

Given a model M , a world w and a formula ϕ , we write $M, w \models \phi$ to mean that ϕ is true at world w in M , under the basic semantics. Truth conditions for atomic formulas, negation and disjunction are entirely standard. The following are truth conditions for the modal operators introduced before.

- $M, w \models \text{After}_{i:\alpha}\phi$ iff $M, w' \models \phi$ for all w' such that $(w', w) \in R_{i:\alpha}$.
- $M, w \models \text{Does}_{i:\alpha}\phi$ iff $\exists w'$ such that $(w', w) \in D_{i:\alpha}$ and $M, w' \models \phi$.
- $M, w \models \text{Bel}_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w', w) \in B_i$.
- $M, w \models \text{Goal}_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w', w) \in G_i$.

The following section is devoted to the semantic properties of \mathcal{L} models and the corresponding axiomatization of our logic \mathcal{L} .

2.1 Properties of the operators

The operators Bel_i , $\text{After}_{i:\alpha}$, $\text{Does}_{i:\alpha}$ and Goal_i are supposed to be normal modal operators satisfying standard axioms and rules of inference of system K . Operators for belief of type Bel_i are supposed to be $KD45$ modal operators, whilst every operator for goal of type Goal_i is supposed to be a KD operator. Thus, we make assumptions about positive and negative introspection for beliefs and we suppose that beliefs and goals cannot be inconsistent.

Actions We add the following constraint over every relation $D_{i:\alpha}$ and $D_{j:\beta}$. For every $i, j \in AGT$ and $\alpha, \beta \in ACT$ and $w \in W$:

$$S1 \quad \text{if } (w', w) \in D_{i:\alpha} \text{ and } (w'', w) \in D_{j:\beta} \text{ then } w' = w'';$$

Constraint $S1$ says that if w' is the *next* world of w which is reachable from w through the occurrence of agent i 's action α and w'' is also the *next* world of w which is reachable from w through the occurrence of agent j 's action β , then w' and w'' denote the same world. Indeed, we suppose here that every world can only have one *next* world. That is, there is an unique next future world after the occurrence of a given action. Property $S1$ corresponds to the following axiom.

$$\text{Alt}_{Act} \text{ Does}_{i:\alpha}\phi \rightarrow \neg \text{Does}_{j:\beta}\neg\phi$$

Axiom **Alt**_{Act} says that: if i is going to do α and ϕ will be true afterward, then it cannot be the case that j is going to do β and $\neg\phi$ will be true afterward.

We also suppose that the world is never static in our framework, that is, we suppose that for every world w there exists some agent i and action α such that i is going to perform α at w . Formally, for every $w \in W$ we have that:

$$S2 \quad \exists i \in AGT, \exists \alpha \in ACT, \exists w' \in W \text{ such that } (w', w) \in D_{i:\alpha}.$$

Property $S2$ corresponds to the following axiom of our logic.

$$\mathbf{Active} \quad \bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha} \top$$

Axiom **Active** ensures that for every world w there is a *next* world of w which is reachable from w by the occurrence of some action of some agent. This is the reason why the operator X for *next* of LTL (linear temporal logic) can be defined as follows.¹

$$X\phi \stackrel{\text{def}}{=} \bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha} \phi$$

The following relationship is supposed between every relation $D_{i:\alpha}$ and the corresponding relation $R_{i:\alpha}$. For every $i \in AGT$, $\alpha \in ACT$ and $w \in W$:

$$S3 \quad \text{if } (w', w) \in D_{i:\alpha} \text{ then } (w', w) \in R_{i:\alpha}.$$

Constraint $S3$ says that if w' is the *next* world of w which is reachable from w through the occurrence of agent i 's action α , then w' is a world which is *possibly* reachable from w through the occurrence of agent i 's action α . Property $S3$ corresponds to the following Axiom **Inc**_{Act}.

$$\mathbf{Inc}_{Act, PAct} \quad Does_{i:\alpha} \phi \rightarrow \neg After_{i:\alpha} \neg \phi$$

According to **Inc**_{Act, PAct}, if i is going to do α and ϕ will be true afterward, then it is not the case that $\neg\phi$ is true after i does α .

Intentions and actions The following axioms relating intentions with actions seem quite natural in the case of intentional actions.

$$\mathbf{IntAct1} \quad (Int_i(\alpha) \wedge Can_i(\alpha)) \rightarrow Does_{i:\alpha} \top$$

$$\mathbf{IntAct2} \quad Does_{i:\alpha} \top \rightarrow Int_i(\alpha)$$

According to **IntAct1**, if i has the intention to do action α and has the capacity to do α , then i is going to do α . According to **IntAct2**, an agent is going to do action α only if he has the intention to do α . In this sense we suppose that an agent's *doing* is by definition intentional.² Similar axioms have been studied in [18, 17] in which a logical model of the relationships between intention and action performance is proposed. **IntAct1** and **IntAct2** correspond to the following semantic constraints over models. For every $i \in AGT$, $\alpha \in ACT$ and $w \in W$:

¹ Note that X satisfies the standard property $X\phi \leftrightarrow \neg X\neg\phi$.

² In our perspective, the category of *doings* (i.e. intentional actions of agents) must be distinguished from the general category of *happenings*. The latter category includes not only physical and natural events, but also spontaneous (non-intentional), atomic behaviors of agents such as reflexes and reactive behaviors of the form *stimulus-response*.

- S4 if $\forall(w', w) \in G_i, \exists w''$ such that $(w'', w') \in D_{i:\alpha}$ and $\exists v$ such that $(v, w) \in R_{i:\alpha}$ then $\exists v'$ such that $(v', w) \in D_{i:\alpha}$;
 S5 if $\exists v'$ such that $(v', w) \in D_{i:\alpha}$ then $\forall(w', w) \in G_i, \exists w''$ such that $(w'', w') \in D_{i:\alpha}$.

Beliefs and goals As far as beliefs and goals are concerned, we only suppose that the two kinds of mental attitudes must be compatible, that is, if an agent has the goal that ϕ he cannot believe that $\neg\phi$. Indeed, the notion of goal we characterize is a notion of an agent's *chosen goal*, i.e. a goal that an agent decides to pursue. As some authors have stressed (e.g. [2]), a rational agent cannot decide to pursue a certain state of affairs ϕ , if he believes that $\neg\phi$ (this has been called *weak realism* hypothesis). Thus, for any $i \in AGT$ and $w \in W$ the following semantic property is supposed:

- S6 $\exists w'$ such that $(w', w) \in B_i$ and $(w', w) \in G_i$.

Property S6 corresponds to the following axiom of our logic.

$$\mathbf{WR} \quad Goal_i\phi \rightarrow \neg Bel_i\neg\phi$$

In this work we assume positive and negative introspection over (chosen) goals, that is, we suppose that the following are axioms of our logic.

$$\mathbf{PIintr} \quad Goal_i\phi \rightarrow Bel_iGoal_i\phi$$

$$\mathbf{NIintr} \quad \neg Goal_i\phi \rightarrow Bel_i\neg Goal_i\phi$$

Axioms **PIintr** and **NIintr** correspond to the following semantic property of models. For any $i \in AGT$ and $w \in W$:

- S7 if $(w', w) \in B_i$ then $\forall v$, if $(v, w) \in G_i$ then $(v, w') \in G_i$;
 S8 if $(w', w) \in B_i$ then $\forall v$, if $(v, w') \in G_i$ then $(v, w) \in G_i$.

Beliefs and actions We suppose that agents satisfy the property of *no forgetting* (NF)³, that is, if an agent i believes that after agent j does α , it is the case that ϕ , and agent i does not believe that j cannot do action α , then after agent j does α , i believes that ϕ . Formally:

$$\mathbf{NF} \quad (Bel_i After_{j:\alpha}\phi \wedge \neg Bel_i \neg Can_j(\alpha)) \rightarrow After_{i:\alpha} Bel_i\phi$$

Axiom **NF** corresponds to the following semantic property of models. For any $i, j \in AGT$, $\alpha \in ACT$, and $w \in W$:

- S9 if $(w', w) \in R_{j:\alpha} \circ B_i$ and $\exists v$ such that $(v, w) \in B_i \circ R_{j:\alpha}$ then $(w', w) \in B_i \circ R_{j:\alpha}$

where \circ is the standard composition operator between two binary relations. In accepting the NF Axiom for beliefs, we suppose that events are always uninformative, that is, i should not forget anything about the particular effects of j 's action α that starts at a

³ See XXX for a discussion of this property.

world w . What an agent i believes at a world w' , only depends on what i believed at the previous world w and on the action which has occurred and which was responsible for the transition from w to w' . Besides, the NF Axiom relies on an additional assumption of complete and correct information. It is supposed that j 's action α occurs if and only if every agent is informed of this fact. Hence all action occurrences are supposed to be public.

We also have specific properties for the actions of informing. We suppose that if an agent i is informed (resp. not informed) by another j about some fact P then i is aware of being informed (resp. not being informed) by j . Formally:

$$\begin{aligned} \mathbf{PBelInf} \quad & Inf_{i,j}(P) \rightarrow Bel_i Inf_{i,j}(P) \\ \mathbf{NBelInf} \quad & \neg Inf_{i,j}(P) \rightarrow Bel_i \neg Inf_{i,j}(P) \end{aligned}$$

Axioms **PBelInf** and **NBelInf** correspond to the following semantic properties of models. For any $i, j \in AGT$, $inf_i(P) \in INFO$, and $w \in W$:

- S10 if $\exists w'$ such that $(w', w) \in D_{j:inf_i(P)}$ then
 $\forall (v, w) \in B_i, \exists w''$ such that $(w'', v) \in D_{j:inf_i(P)}$
- S11 if $\forall (v, w) \in B_i, \exists w''$ such that $(w'', v) \in D_{j:inf_i(P)}$ then
 $\exists w'$ such that $(w', w) \in D_{j:inf_i(P)}$

We call \mathcal{L} the logic axiomatized by the axioms and rules of inference presented above. We write $\vdash \varphi$ if formula φ is a Theorem of \mathcal{L} . We write $\models \varphi$ if φ is *valid* in all \mathcal{L} models, i.e. $M, w \models \varphi$ for every \mathcal{L} model M and world w in M . Finally, we say that φ is *satisfiable* if there exists a \mathcal{L} model M and world w in M such that $M, w \models \varphi$.

3 A general definition of trust

In this work trust is conceived as a complex configuration of mental states in which there is both a motivational component and a doxastic component. More precisely, we assume that i 's trust in agent j necessarily involves a goal of the truster: if i trusts agent j then necessarily i trusts j with respect to some of his goals. The core of trust is a belief of the truster about some properties of the trustee, that is, if i trusts agent j then necessarily i trusts j because i has some goal and believes that j has the right properties to ensure that such a goal will be achieved. The concept of trust formalized in this work is similar to the concept of trust defined by Castelfranchi & Falcone [4]. We agree with them that trust should not be seen as an unitary and simplistic notion as other models implicitly suppose. For instance, there are computational models of trust in which trust is conceived as an expectation of the truster about a successful performance of the trustee sustained by the repeated direct interactions with the trustee (under the assumption that iterated experiences of success strengthen the truster's confidence) [15, 21]. More sophisticated models of social trust have been developed in which reputational information is added to information obtained via direct interaction (e.g. [12, 20]). All these models are in our view over-simplified since they do not consider the beliefs supporting the truster's expectation which enter into play in the truster's evaluation of the trustee. On this point we agree with Castelfranchi & Falcone on the fact that: trust is based on the truster's ascription of specific properties to the trustee (e.g. abilities,

competencies, dispositions, etc.) and to the environment in which the trustee is going to act which are relevant for the achievement of a goal of the truster.

Here we just focus on a particular form of trust that can be called *trust in the trustee's actions*. According to the proposed definition, agent i trusts agent j to do a certain action α if and only if i has a certain goal and thinks that j will perform action α in such a way that his goal will be achieved. In a complementary work we have provided a richer typology of trust and distinguished *trust in the trustee's actions* from *trust in the trustee's dispositions* (or *trust in the trustee's willingness*). According to the proposed definition, agent i trusts agent j to be willing to do a certain action α for him if and only if i has a certain goal and thinks that j is willing to perform action α for him in such a way that his goal will be achieved. We defined the concept of willingness as follows: j is willing to do action α for i if and only if j has the conditional goal to perform action α under the condition in which he believes that i wants him to do α . The notion of *trust in the trustee's dispositions* will be ignored in the present paper.

We first define a notion of positive trust in the trustee's action.

Definition 1 POSITIVE TRUST ABOUT ACTION. i trusts j to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:⁴

1. j , by doing α , will ensure that ϕ AND
2. j has the capacity to do α AND
3. j intends to do α

The three conditions 1, 2 and 3 can reformulated in formal terms as follows.

1. Condition C1: $After_{j:\alpha}\phi$
2. Condition C2: $Can_j(\alpha)$
3. Condition C3: $Int_j(\alpha)$

Condition C1 concerns the trustee's power to satisfy the truster's goal that ϕ by means of the performance of action α . Conditions C2 and C3 are about the trustee's properties which are necessary and sufficient for him to perform action α . The formal translation of Definition 1 is:

$$Trust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X \phi \wedge Bel_i (After_{j:\alpha}\phi \wedge Can_j(\alpha) \wedge Int_j(\alpha))$$

In our logic conditions C2 and C3 together are equivalent to $Does_{j:\alpha}\top$ (by Axiom **IntAct2**), so the definition of trust can be simplified as follows:

$$Trust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X \phi \wedge Bel_i (After_{j:\alpha}\phi \wedge Does_{j:\alpha}\top)$$

$Trust(i, j, \alpha, \phi)$ is meant to stand for: i trusts j to do α with regard to his goal that ϕ .

⁴ In the present paper we only focus on *full trust* involving a *certain belief* of the truster. In order to extend the present analysis to forms of *partial trust*, a notion of *graded belief* (i.e. uncertain belief) or *graded trust*, as in [7], is needed.

Example 1. Suppose that Bill trusts Mary to shoot Bob with regard to his goal that Bill will die in the next state:

$$Trust(Bill, Mary, shoot, \neg BobAlive).$$

This means that Bill wants Bob to die in the next state:

$$Goal_{Bill} X \neg BobAlive.$$

Moreover, according to Bill's beliefs, Mary, by shooting Bob, will ensure that Bob is dead in the next state, and Mary is going to shoot Bob:

$$Bel_{Bill}(After_{Mary:shoot} \neg BobAlive \wedge Does_{Mary:shoot} \top).$$

The following theorem highlights the fact that if i trusts j to do α with regard to his goal that ϕ then i has a positive expectation that ϕ will be true in the next state.

Theorem 1 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

$$\vdash Trust(i, j, \alpha, \phi) \rightarrow Bel_i X \phi$$

The dual notion of negative trust in the trustee's action is based on the fact that, by doing some action α , agent j can prevent i to reach his goal. In that case i expects that j will not intend to do α .

Definition 2 NEGATIVE TRUST ABOUT ACTION. *i trusts j not to do α with regard to his goal ϕ if and only if i wants ϕ to be true and i believes that:*

1. j , by doing α , will ensure that $\neg\phi$ AND
2. j does not intend to do α

The two conditions 1 and 2 can be reformulated in more formal terms as follows.

1. Condition C1: $After_{j:\alpha} \neg\phi$
2. Condition C2: $\neg Int_j(\alpha)$

The formal definition of negative trust in the trustee's action is given by the following abbreviation.

$$Trust(i, j, \neg\alpha, \phi) \stackrel{\text{def}}{=} Goal_i X \phi \wedge Bel_i (After_{j:\alpha} \neg\phi \wedge \neg Int_j(\alpha))$$

$Trust(i, j, \neg\alpha, \phi)$ stands for: i trusts j to do α with regard to his goal that ϕ .⁵

Example 2. Suppose that Bill trusts Mary not to shoot him with regard to his goal to be alive in the next state:

$$Trust(Bill, Mary, \neg shoot, BobAlive).$$

This means that Bill wants to be alive in the next state:

$$Goal_{Bill} X BillAlive.$$

Moreover, according to Bill's beliefs, Mary, by shooting him, will ensure that he will be dead in the next state, but she does not intend to shoot him:

$$Bel_{Bill}(After_{Mary:shoot} \neg BillAlive \wedge \neg Int_{Mary}(shoot)).$$

In this sense, Bill's (negative) trust in Mary is based on Bill's belief that Mary is in condition to kill him by shooting him, but she does not have the intention to shoot (i.e. she does not intend to perform the action which will frustrate Bill's goal).

⁵ Note that $Trust(i, j, \neg\alpha, \phi) \wedge \neg Bel_i X \phi$ is satisfiable in our logic, that is, i 's negative trust in j 's action α with regard to i 's goal that ϕ does not entail i 's positive expectation that ϕ will be true. The intuitive reason is that $\neg\phi$ may be the effect of another action than $j : \alpha$.

4 Properties of an information source

In previous works [9, 7, 10] we have provided a logical characterization of several properties of information sources such as sincerity, competence, vigilance, credibility, etc. For example, agent j is said to be credible with respect to the diagnosis of a certain disease if and only if if j says that someone has this disease, then it is the case that he has it. The properties of an information source j can be defined depending on what the information source j tells to agent i , what the information source j believes, and what is really the case. This means that each property can be defined in terms of the relationships between the formal constructions: Bel_j , $inf_i(P)$, and P . For instance, if after j informs i about P it is the case that j believes P , then j is said to be sincere about P with regard to agent i . Here we just focus on some of these properties.

Definition 3 VALIDITY OF AN INFORMATION SOURCE. Agent j is a valid information source about P with regard to i if and only if after j does the action of informing i about P , it is the case that P .

Formally:

$$Valid(j, i, P) \stackrel{\text{def}}{=} After_{j:inf_i(P)} P$$

Definition 4 COMPLETENESS OF AN INFORMATION SOURCE. Agent j is a complete information source about P with regard to i iff if P is true, then j does the action of informing i about P .

Formally:

$$Compl(j, i, P) \stackrel{\text{def}}{=} P \rightarrow Does_{j:inf_i(P)} \top$$

Definition 5 SINCERITY OF AN INFORMATION SOURCE. Agent j is a sincere information source about P with regard to i iff after j does the action of informing i about P , j believes that P .

Formally:

$$Sinc(j, i, P) \stackrel{\text{def}}{=} After_{j:inf_i(P)} Bel_j P$$

Definition 6 COMPETENCE OF AN INFORMATION SOURCE. Agent j is a competent information source about P iff if j believes P , then P is true.

Formally:

$$Compet(j, P) \stackrel{\text{def}}{=} Bel_j P \rightarrow P$$

Definition 7 VIGILANCE OF AN INFORMATION SOURCE. Agent j is a vigilant information source about P iff if P is true, then j believes P .

Formally:

$$Vigil(j, P) \stackrel{\text{def}}{=} P \rightarrow Bel_j P$$

Definition 8 COOPERATIVITY OF AN INFORMATION SOURCE. Agent j is a cooperative information source about P with regard to i iff if j believes P , then j does the action of informing i about P .

Formally:

$$Coop(j, i, P) \stackrel{\text{def}}{=} Bel_j P \rightarrow Does_{j:inf_i(P)} \top$$

The previous properties of information sources are not independent. For instance, as the following Theorem 2 shows, sincerity and competence entail validity.

Theorem 2 *Let $i, j \in AGT$ and $inf_i(P) \in INFO$, then:*

$$\vdash (Sinc(j, i, P) \wedge After_{j:inf_i(P)} Compet(j, P)) \rightarrow Valid(j, i, P)$$

Theorem 2 says that if j is a sincere information source about P with regard to i and after j informs i about P , j is a competent information source about P , then j is a valid information source about P with regard to i .

Example 3. Let's consider an example in the field of stocks and bonds market where trust plays a significant role. The agent BUG is the Bank of Union of Groenland. Sue Naive (S.N.) and Very Wise (V.W.) are two BUG's customers. BUG plays the role of an information source for the customers, for instance for the propositions:

P : "it is recommended to buy MicroHard stoks".

Q : "Microhard stocks are dropping".

S.N. believes that BUG is sincere with regard to her about P and BUG is competent about P , because she believes that BUG wants to help his customers and BUG has a long experience in the domain. She also believes that BUG is cooperative with regard to her about Q because Q is a relevant information for customers to take decisions.

V.W. also believes that BUG is competent about P , but he does not believe that BUG is sincere with regard to him about P , and that BUG is not cooperative with regard to him about Q , because V.W. believes that BUG wants that V.W. buy Microhard stocks, even if this is not profitable for the customers.

According to our definitions, this example is formally represented by:

$$\begin{aligned} & Bel_{S.N.}(Sinc(BUG, S.N., P) \wedge Compet(BUG, P) \wedge Coop(BUG, S.N., Q)) \\ & Bel_{V.W.}(Compet(BUG, P)) \wedge \neg Bel_{V.W.}(Sinc(BUG, V.W., P) \wedge \\ & \neg Bel_{V.W.}(Coop(BUG, V.W., Q)) \end{aligned}$$

5 Trust in information sources

We conceive trust in information sources as a specific instance of the general notion of trust as defined in section 3. In our view, the relevant aspect of trust in information sources is the content of the truster's goal. In particular, we suppose that an agent i trusts the information source j to inform him about the truth value of proposition p only if i has the *epistemic goal* of knowing whether p is true and believes that thank to the information transmitted by j he will achieve this goal. In this sense, trust in information sources is characterized by an epistemic goal of the truster and an informative action of the trustee.

Thus, in order to provide a formal analysis of trust in information sources, we need to introduce a concept of epistemic goal. An *epistemic goal* of an agent i is a goal of i of knowing the truth value of a certain proposition. The definition of *epistemic goal*

is built on the basis of the following two standard definitions of *knowing that* (i.e. as having the correct belief that something is the case) and *knowing whether*:

$$K_i\phi \stackrel{\text{def}}{=} Bel_i\phi \wedge \phi$$

$$KW_i\phi \stackrel{\text{def}}{=} K_i\phi \vee K_i\neg\phi$$

where $K_i\phi$ stands for: agent i knows that ϕ is true,⁶; $KW_i\phi$ stands for: i knows whether ϕ is true.

Our aim in this section of the paper is to investigate the relationships between trust in information sources and the properties of information sources defined in section 4. The following Theorem 3.1 highlights the relationship between trust in information sources and the properties of validity and completeness of information sources. It says that: if i believes that j is a valid information source about p and $\neg p$ with regard to i and that j is a complete information source about p and $\neg p$ with regard to i , i believes that j can inform him about p and can inform him about $\neg p$, and i has the epistemic goal of knowing whether p , then i trust the information source j to inform him about p or i trust the information source j to inform him about $\neg p$ with regard to his epistemic goal of knowing whether p . Theorem 3.2 says that: if i trust the information source j to inform him about p or i trust the information source j to inform him about $\neg p$ with regard to his epistemic goal of knowing whether p , then i believes that in the next state he will achieve his epistemic goal of knowing whether p .

Theorem 3 *Let $i, j \in AGT$ and $inf_i(p), inf_i(\neg p) \in INFO$, then:*

1. $\vdash (Bel_i(Valid(j, i, p) \wedge Valid(j, i, \neg p) \wedge Compl(j, i, p) \wedge Compet(j, i, \neg p)) \wedge Goal_iKW_i p \wedge Bel_i(Can_j(inf_i(p)) \wedge Can_j(inf_i(\neg p)))) \rightarrow Trust(i, j, inf_i(p), KW_i p) \vee Trust(i, j, inf_i(\neg p), KW_i p)$
2. $\vdash Trust(i, j, inf_i(p), KW_i p) \vee Trust(i, j, inf_i(\neg p), KW_i p) \rightarrow Bel_iXKW_i p$

Theorem 4 *Let $i, j \in AGT$, then:*

1. $\vdash Bel_i(Sinc(j, i, p) \wedge Sinc(j, i, \neg p) \wedge Compl(j, i, p) \wedge Compet(j, i, \neg p)) \wedge Goal_iKW_i Bel_j p \wedge Bel_i(Can_j(inf_i(p)) \wedge Can_j(inf_i(\neg p))) \rightarrow Trust(i, j, inf_i(p), KW_i Bel_j p) \vee Trust(i, j, inf_i(\neg p), KW_i Bel_j p)$
2. $\vdash Trust(i, j, inf_i(p), KW_i Bel_j p) \vee Trust(i, j, inf_i(\neg p), KW_i Bel_j p) \rightarrow Bel_iXKW_i Bel_j p$

The reason why in Theorem 4 we need the hypothesis $Bel_i(Compl(j, i, p) \wedge Compet(j, i, \neg p))$ instead of $Bel_i(Coop(j, i, p) \wedge Coop(j, i, \neg p))$ is that $Bel_j p \vee Bel_j \neg p$ is not a theorem. Then, we cannot infer $Bel_i Int_j(inf_i(p)) \vee Bel_i Int_j(inf_i(\neg p))$ from $Bel_i(Coop(j, i, p) \wedge Coop(j, i, \neg p))$.

⁶ This definition of knowledge is purely technical. A more intuitive definition of “agent i knows that ϕ ” is “agent i has the justified belief that ϕ ”.

6 Related works on logic of trust

As far as logics of trust are concerned, we think that there is still no adequate logical characterization of this concept. Indeed, most of logical models of trust have been exclusively focused on trust in information and communication sources [16, 14, 8, 6]. Other models have reduced trust to a kind of belief or expectation of the truster ignoring the motivational aspect of trust. For instance, in [7] trust is reduced to a kind of belief of the truster (called “strong belief”) about a particular property of the trustee (e.g. sincerity, credibility, cooperativity). In [13] trust is characterized on the basis of two kinds of beliefs of the truster: the truster’s belief that a certain rule or regularity applies to the trustee (called “rule belief”), and the truster’s belief that the rule or regularity is going to be followed by the trustee (called “conformity belief”). From this perspective there is a pressing need for elaborating logical models of trust in which the truster’s beliefs about the relevant properties of the trustee (power, abilities, motivational dispositions, moral dispositions, etc.) are adequately modeled, and in which the motivational aspect of trust is taken into account.

7 Conclusion

We have presented in a modal logical framework a model that integrates in trust definition: the truster’s motivation (his goal), the action that allows the trustee to reach this goal, and the trustee’s ability and intention to do this action. It has been shown in Theorem 1 that if some assumptions about these properties hold, it follows that the truster believes that his goal will be reached in the next state. An original feature of the model is to show that the truster may expect that the trustee does, or does not do, the action.

In the same framework we have defined expected trustee’s epistemic properties: sincerity, competence, vigilance, cooperativity, validity and completeness, and the relationships between these properties are given in Theorem 2. Finally, Theorems 3 and 4 show that, under some assumptions expressed in terms of these properties, the truster trusts that the appropriate informative action allows the truster to reach his epistemic goal.

These results could be extended in further works by a deeper analysis of the temporal dimension. In particular there are many situation where the truster does not have a goal in the actual situation, but he believes that he may have this goal in some future situations, and he trusts the trustee for reaching his goal in these future situations. For instance, a customer may trust a consultant on being informed if the value of a given stock drops under a fixed threshold. From a technical point of view, this extension needs to add more general temporal modalities than the “next” operator, but that would not require to change the foundations of our analysis of trust.

References

1. L. Åqvist. Deontic logic. In D. M. Gabbay and F. Geunther, editors, *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht, 2002.
2. M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, 1987.

3. J. Carmo and A. Jones. Deontic Logic and Contrary-to-Duties. In D. Gabbay, editor, *Handbook of Philosophical Logic (Rev. Edition)*. Reidel, to appear.
4. C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
5. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
6. M. Dastani, A. Herzig, J. Hulstijn, and L. van der Torre. Inferring trust. In *Proc. of Fifth Workshop on Computational Logic in Multi-agent Systems (CLIMA V)*, volume LNCS 3487, pages 144–160. Springer, 2004.
7. R. Demolombe. Reasoning about trust: a formal logical framework. In *Trust management: Second International Conference iTrust*, volume LNCS 2995, pages 291–303. Springer.
8. R. Demolombe. To trust information sources: a proposal for a modal logical framework. In C. Castelfranchi and Y.-H. Tan, editor, *Trust and Deception in Virtual Societies*. Kluwer, 1999.
9. R. Demolombe. To trust information sources: a proposal for a modal logical framework. In C. Castelfranchi and Y.-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.
10. R. Demolombe and C.-J. Liao. A logic of graded trust and belief fusion. In *Proc. of Fourth Workshop on Deception, Fraud and Trust in Agent Societies*, pages 13–25, 2001.
11. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.
12. T. G. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agent and Multi-Agent Systems*, 13:119–154, 2006.
13. A. J. I. Jones. On the concept of trust. *Decision Support Systems*, 33(3):225–232, 2002.
14. A. J. I. Jones and B. S. Firozabadi. On the characterization of a trusting agent: Aspects of a formal approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
15. C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In *Multi-Agent System Engineering: Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*, pages 221–231. Springer, 1999.
16. C. J. Liao. Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.
17. E. Lorini. *Variations on intentional themes: From the generation of an intention to the execution of an intentional action*. PhD thesis, University of Siena, Department of Philosophy, 2007.
18. E. Lorini, A. Herzig, and C. Castelfranchi. Introducing “attempt” in a modal logic of intentional action. In *Logics in Artificial Intelligence: 10th European Conference (JELIA 2006)*, volume LNAI 4160, pages 280–292. Springer, 2006.
19. J.-J. C. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
20. J. Sabater and C. Sierra. Regret: a reputation model for gregarious societies. In *Proc. of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 475–482, 2001.
21. M. Witkowski, A. Artikis, and J. Pitt. Experiments in building experiential trust in a society of objective-trust based agents. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 111–132. Kluwer Academic Publishers, Dordrecht, 2001.